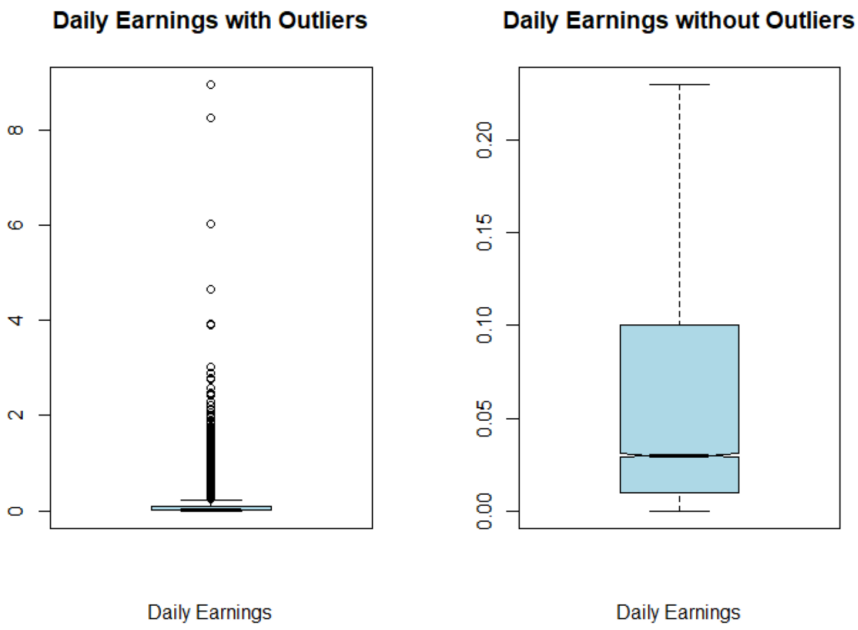
Project #3: YouTube Attention and Review

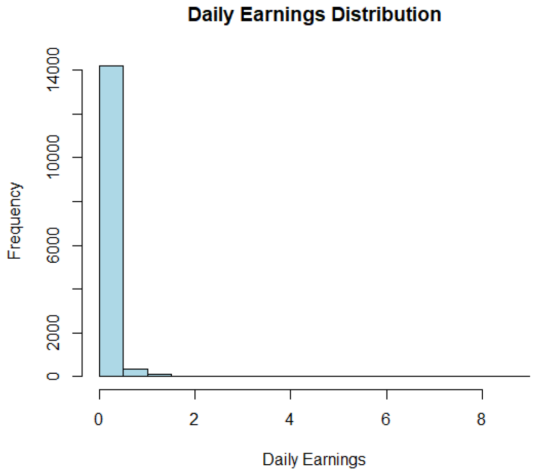
Group #7
Jaydip Patel, Vincent Chang, Nicole Haberer, Ivy Wang, Elise Lu, & Griffin Parr

Part 1:

Daily Earnings Summary

Mean: 0.1008
Standard Deviation: 0.2338
Median: 0.0300

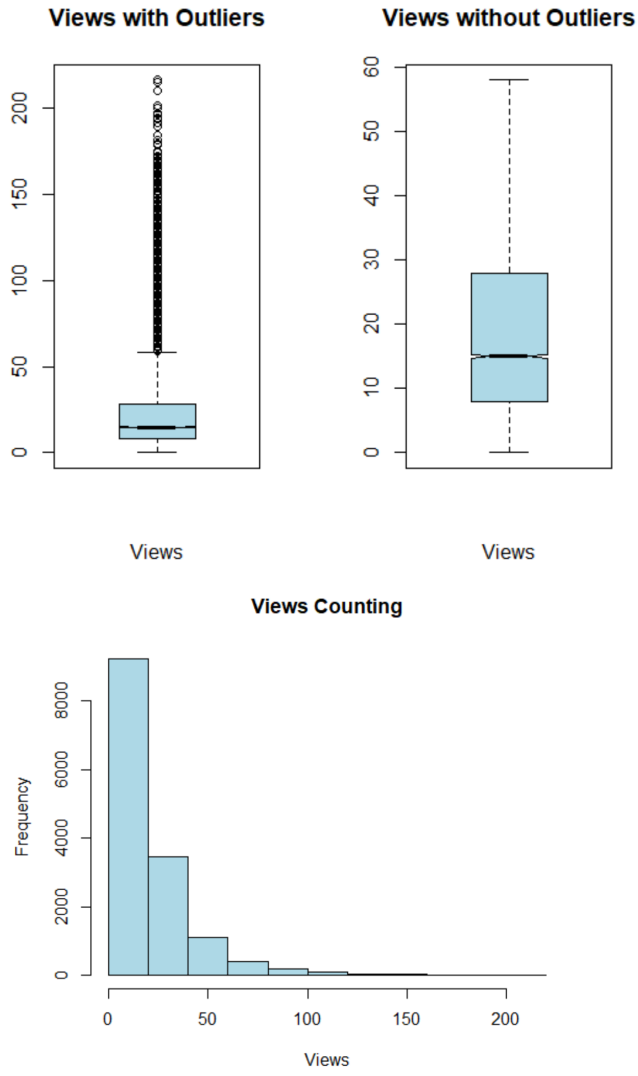




Daily Earnings shows a few significant outliers, but most of the earnings are under $1/day. The significant skew in this data indicates that most of our income can be attributed to a small number of videos. These outliers do not seem to be typos, but relevant pieces of data that tells us about the skewed nature of our profitability.

Views Summary

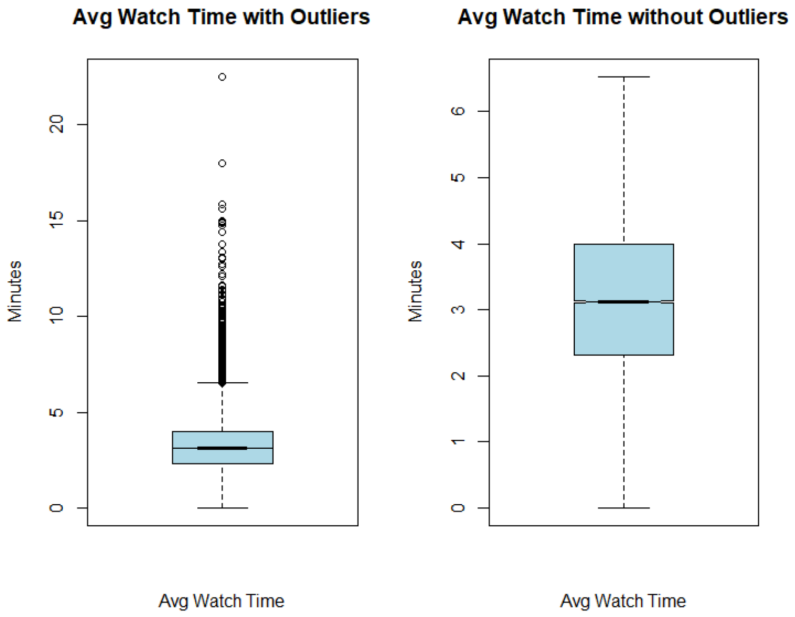
Mean: 22.3
Standard Deviation: 22.97595
Median: 15.0

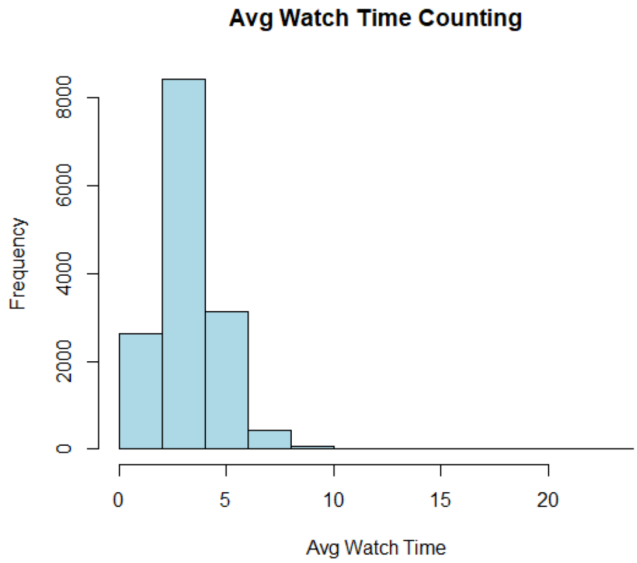


Views are also highly skewed with a long right tail. Notice the heavy right tail have significant impact on the mean as the median is smaller than the average, with the great dispersion between the number of views. Despite the indication of these outliers, we wouldn't consider removing any data points in this case, because these are not outliers, but an indicator that a few videos are receiving a significantly larger proportion of the views.

Average Watch Time Summary

Mean: 3.232
Standard Deviation: 1.47565
Median: 3.120





Average watch time is fairly normally distributed with a mean of 3.2 and a slight right tail. It is centered around a mean of 3.2 indicating a fairly short attention span for our viewers.

Part 2:

#1

During the pre-period (before Feb 5th):

- Is the average watch time different for videos that have pre-roll ads versus not?
    - H0 = The average watch time is not different with pre-rolls ads versus those that do not have pre-roll ads
    - H1 = The average watch time is different for videos with pre-rolls ads versus those that do not have pre-roll ads

```
        Welch Two Sample t-test

data:  Q1_pre$average_watch_time and Q1_no$average_watch_time
t = -8.3977, df = 11217, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2792857 -0.1735792
sample estimates:
mean of x mean of y
 3.104515  3.330947
```

This t-test tells us that we should reject the null hypothesis because the average watch time difference between videos with pre-roll ads and those without pre-roll is statistically significant. With a p-value of less than .001 and a t-test statistic of -8.3977, which translates to the mean of pre-roll average watch time being over 8 standard errors below the dataset mean of no pre-roll average watch time.

- Is the number of views different for videos that have pre-roll ads versus not?
    - H0 = The number of views is not different for videos that have pre-roll ads versus not
    - H1 = The number of views is different for videos that have pre-roll ads versus not

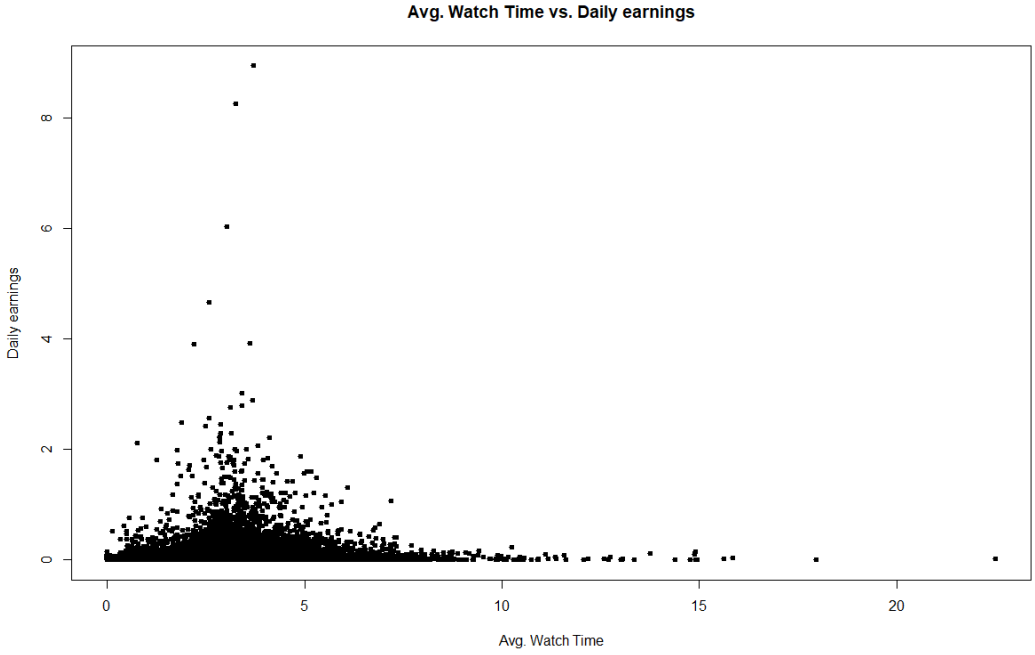
```
        Welch Two Sample t-test

data:  Q1_pre$Views and Q1_no$Views
t = -12.159, df = 9896, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.006455 -4.338688
sample estimates:
mean of x mean of y
 20.12362  25.29620
```

This t-test tells us that we should reject the null hypothesis because the views for videos with pre-roll ads and those without is statistically significant. With a p-value of less than .05 and a t-test statistic of -12.159, we see that the mean of pre-roll views is more than 12 standard errors below the dataset mean of no pre-roll views. Notice that with ads, views and average watch time is more spread out and significantly more right skewed.

#2
For the full sample:
Effect of Average Watch Time on Daily Earnings
Daily_earnings = 0.082214 + 0.005762*(average_watch_time)

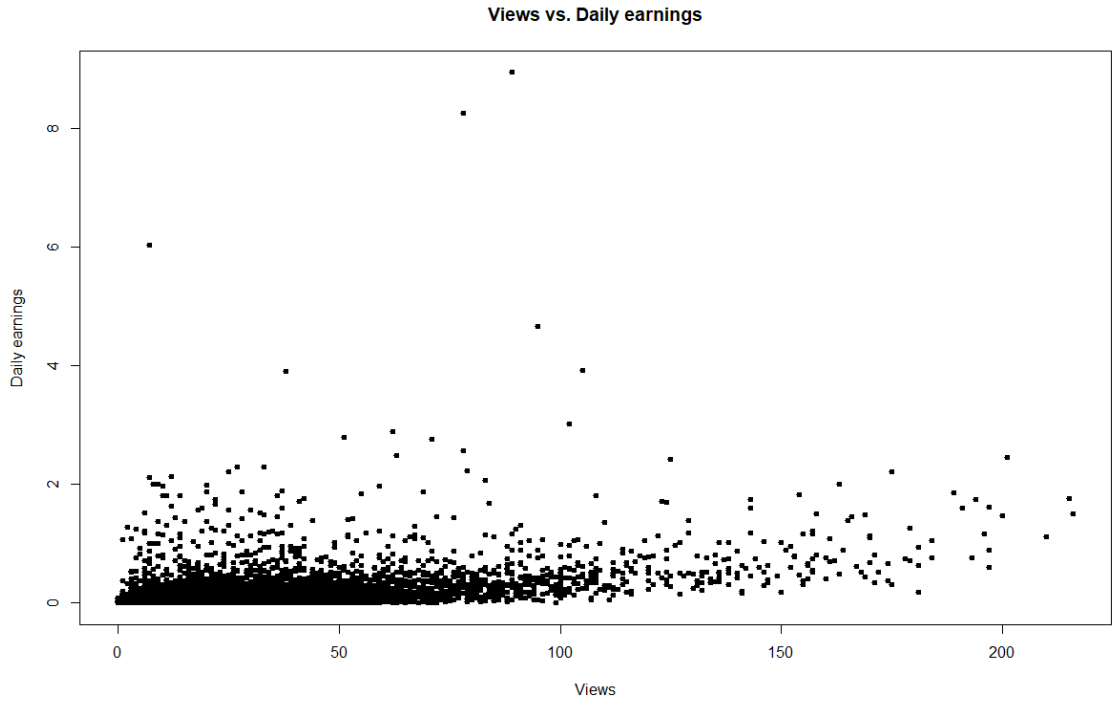
The relationship between these two variables is statistically significant as p-value of the slope is less than 0.05. For every one unit increase in average watch time, there is a 0.00576 dollar increase in daily earning. The intercept is statistically significant with a p-value less than 0.05. The intercept shows that at a watch time of 0 (x=0), daily earnings is predicted to be $0.0822. To us, this indicates that this intercept is meaningless or that youtubers may receive a base amount of compensation for simply uploading a video to their channel. In addition, a R-square of 0.125% indicating that 0.125% of variability in daily earnings can be explained by average watch time.



Effect of Views on Daily Earnings

Daily_earnings = -3.074e-03 + 4.659e-03*(views)

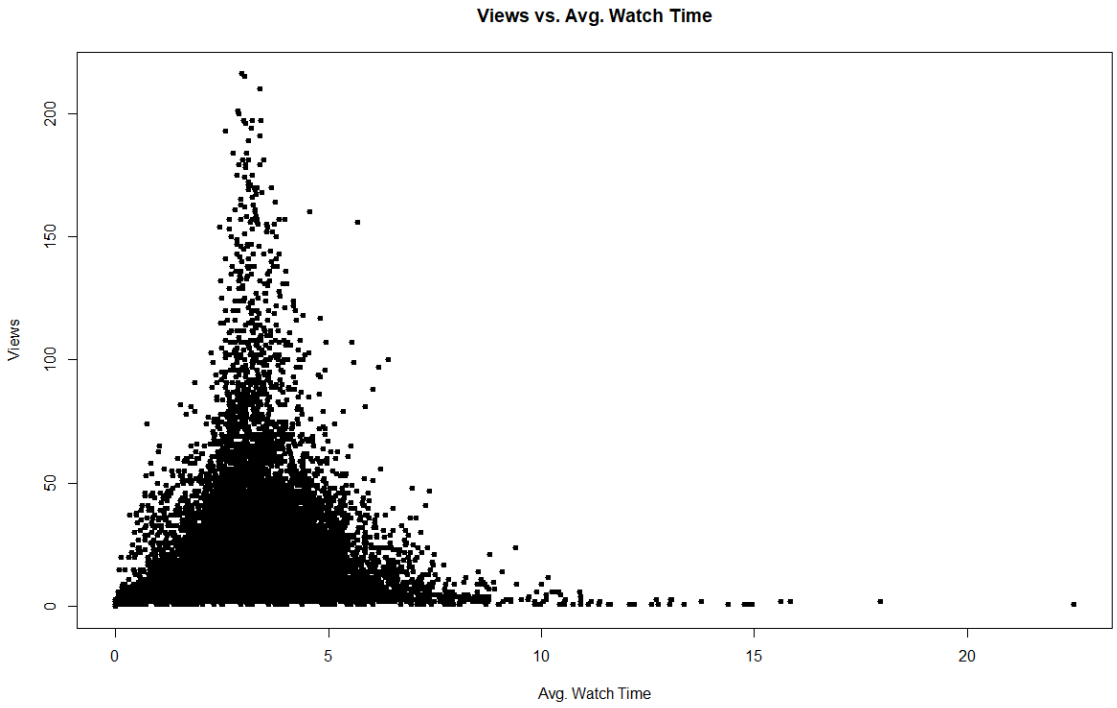
The relationship between these two variables is statistically significant because the p-value is less than 0.05. For every unit increase in views, there is a 4.659e-03 unit increase in daily earnings. The intercept is not statistically significant and is also meaningless because it impossible to have negative daily earnings (y=-3.074e-03 at x=0). In addition, a R-square of .2095 indicates that 20.95% of variability in daily earnings can be explained by views.



Effect of Average Watch Time on Views

Views = 20.8245 + 0.4572*(average_watch_time)

The relationship between these two variables is statistically significant because the p-value is less than 0.05. For every one unit increase in average watch time, there is a 0.4572 unit increase in views. The intercept is statistically significant and meaningful with p-value less than 0.05. For an average watch time of 0 (x=0), there will be 20.82 views on average. This make sense because it is fairly common to click on a video by mistake and then quickly leave the video, registering a view but a watch time of zero. In addition, a R-square of .000794 indicates that 0.0794% of the variability in views can be explained by average watch time.



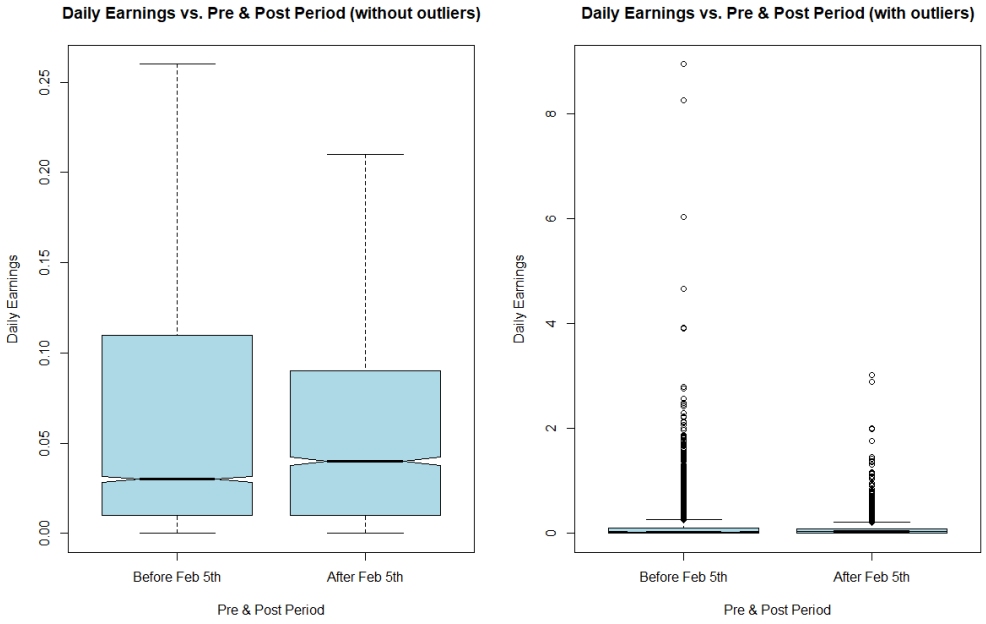
#3

Effect of Post on Daily Earnings

Daily_earnings = 0.104190 - 0.022166*(post)

The impact of post on daily earnings is significant with a p-value less than 0.05, indicating that there is a difference in daily earnings between pre and post period. The statistically significant intercept shows that the average daily earning before February 5th is 0.104, while the average daily earnings decrease by 0.0221 after February 5th. This decrease in daily earnings can be attributed to significant outliers in daily earning before February 5th which significantly increased the overall average earnings. R-squared of .00144 indicates that 0.144% of variability in daily earnings can be explained by pre/ post-period.

In addition, the great difference in the sample size of pre (n=11371) and post period (n=3097) may have unpredictable impact the linear regression model.

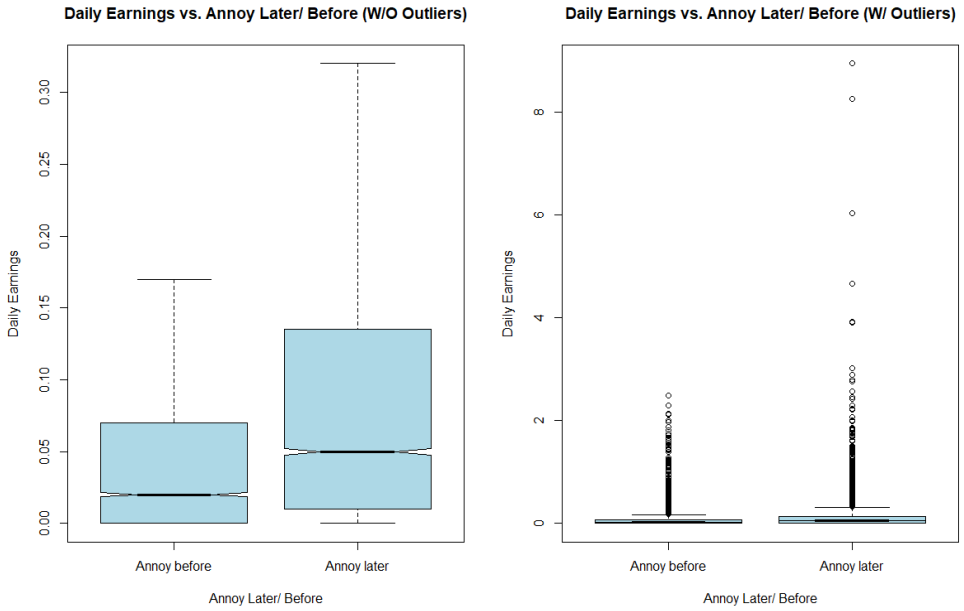


#4

Effect of Annoy Later on Daily Earnings

daily earnings = 0.068878 + 0.055371*(annoy_later)

The impact of annoy_later on daily earnings was significant with a p-value of less than 0.05, indicating that there is a difference in daily earnings when videos went from no pre-roll advertisements to pre-roll advertisements and pre-roll advertisements to no pre-roll advertisements during the post period. In the other words, on average, the videos that were selected to run ads before Feb 5th are less profitable than those after Feb 5th. The statistically significant intercept shows that the average daily earning for videos chosen to pre-roll ads in pre-period is 0.0688, while the average daily earnings increase by 0.0553 for those have pre-roll ads in post-period. In addition, R-squared of .01.37 indicates that 1.37% of variability in daily earnings can be explained by the time period of when the ads were shown (annoy later or before).



#5

Effect of Post and Annoy Later (and interaction) on Daily Earnings

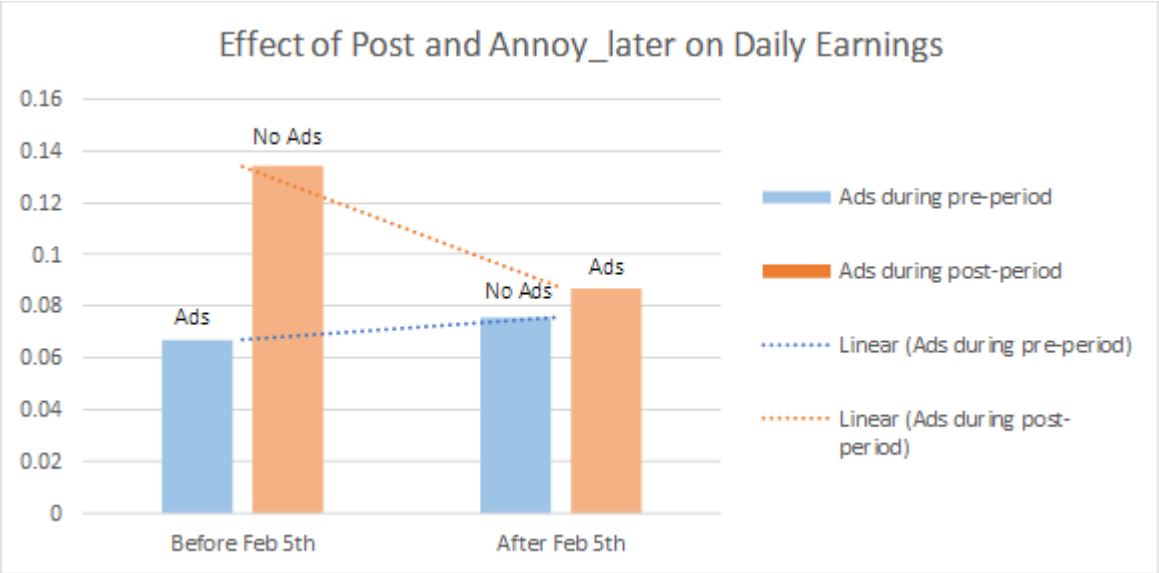
Daily_Earnings = 0.066961 + 0.008918(post)+0.067377*(annoy_later)-0.056208*(post)*(annoy_later)

With the multiple regression model, the main effect of annoy later is statistically significant ($p < 0.05$) while the main effect of post is not. This indicates that, on average, the videos that were selected to run ads before Feb 5th make $0.0673 less than those run after Feb 5th. In previous

linear regression models such as #3, the relationship shows that pre-period generates more profit than post-period, but in this model the effect of post is statistically insignificant so there is no real difference in daily earnings between videos before or after Feb 5th. However, once we combine post with annoy later, they jointly provide an explanation for daily earnings even if we do not have the statistical power to disentangle their individual effects here.

An R-squared of .0175, indicates that the two variables, annoy later and post together can explain 1.75% of the variability in daily earnings. In summary, videos without ads before Feb 5th generate highest daily earnings ($0.134), followed by videos with ads after Feb 5th ($0.0870), followed by videos with ads after Feb 5th ($0.0758) and finally without ads before Feb 5th ($0.0669).

Daily Earnings	**Post** = 0 (Before feb 5th)	**Post** = 1 (After feb 5th)	Sum across the row
Annoy_later = 0 (Ads during pre-period)	0.0669 (Ads)	0.0758 (No Ads)	0.1427
Annoy_later = 1 (Ads during post-period)	0.134 (No Ads)	0.0870 (Ads)	0.221
Sum across the column	0.201	0.162	



#6

Effect of Post on Average Watch time
Average Watch Time = 3.22963 + 0.11862*(post)

The impact of post on average watch time was significant with a p-value less than 0.05, indicating that there is a difference in average watch time between the pre and post periods. The statistically significant intercept shows that the average watch time before February 5th is 3.22, while the average watch time increased by 0.11862 after February 5th. This might indicate that as time passed, the channel started to grab more attention. An R-squared of .00103 indicates that 0.103% of variability in average watch time can be explained by timing.

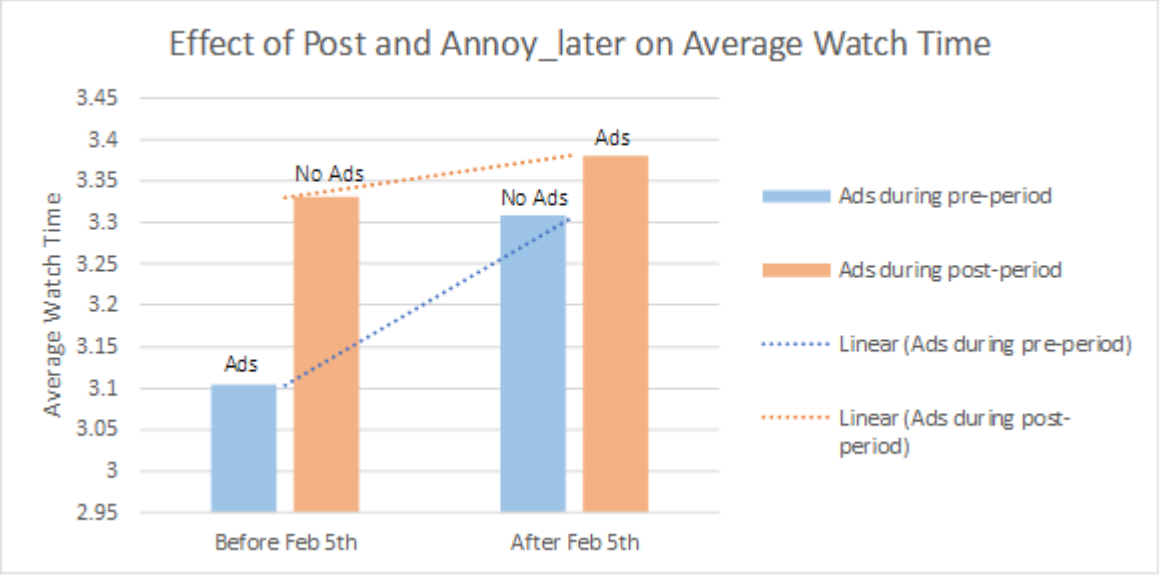
Effect of Annoy Later on Average Watch time
average_watch_time = 3.14827 + 0.19337*(annoy_later)

The impact of annoy_later on average watch time was significant with a p-value less than 0.05, indicating that there is a difference in average watch time between videos that transitioned from no pre-roll advertisements to pre-roll advertisements and vice versa during the post period. This means that on average, the videos that were selected to run ads before Feb 5th attract more attention than those after Feb 5th. The intercept shows that the average watch time for videos chosen to run pre-roll ads in the pre-period is 3.14, while the average watch time increases by 0.193 for those that have pre-roll ads in post-period. In addition, an R-squared of .00422 indicates that 0.422% of variability in average watch time can be explained by annoy later or before.

Effect of Post and Annoy Later (and interaction) on Average Watch Time
Average_watch_time =3.10451 +0.20358*(post) + 0.22643*(annoy_later) - 0.15345*(post)*(annoy_later)

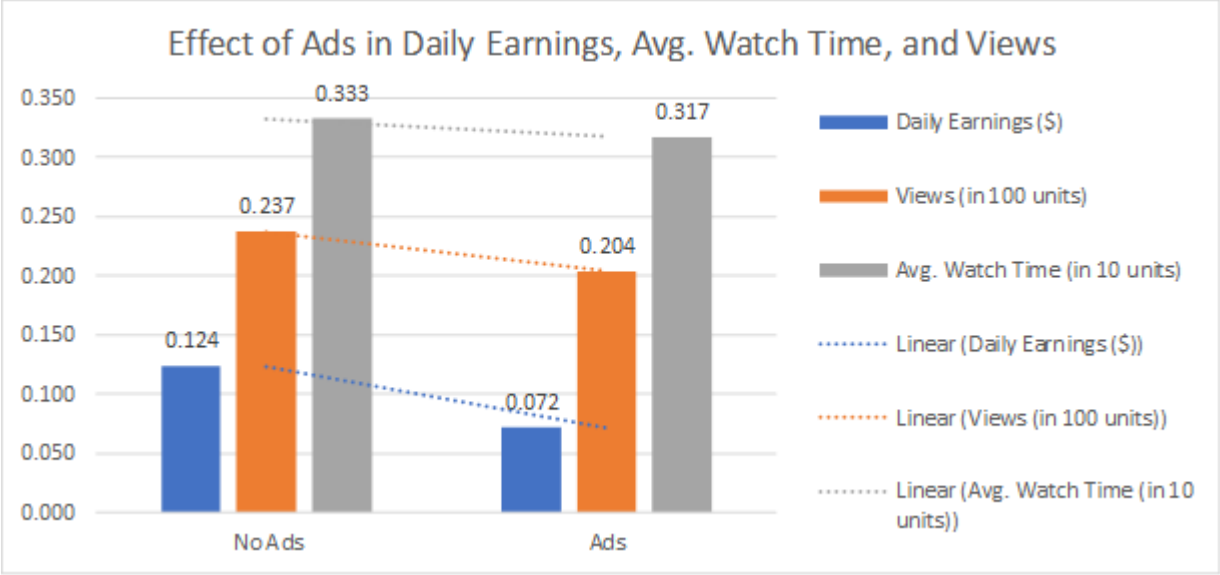
With the multiple regression model, all important variables are statistically significant ($p < 0.05$), including the interaction between annoy later and post. While the main effect of post and annoy later preserved the pattern observed in previous linear regression model, this model helps us identify the best and worst scenarios for achieving the highest watch time. Videos with ads after Feb 5th attracted most attention (3.381), follow by videos without ads before Feb 5th (3.33)and the videos without ads after Feb 5th (3.308), and finally, the videos with ads before Feb 5th attracted the least attention (3.104).

Average Watch Time	**Post** = 0 (Before feb 5th)	**Post** = 1 (After feb 5th)	Sum across the row
Annoy_later = 0 (Ads during pre-period)	3.104 (Ads)	3.308 (No Ads)	6.41
Annoy_later = 1 (Ads during post-period)	3.33 (No Ads)	3.381 (Ads)	7.14
Sum across the column	6.43	6.68	



Conclusion:

After reviewing the data, we believe we have come to several conclusions that a YouTuber can use to guide his or her decisions regarding pre-roll ads running on a channel. In general for all videos, we determined that introducing pre-roll ads when none were present decreases daily earnings, views, and average watch time (this difference is statistically significant across all variables). The magnitude of this impact varies based on other factors such as annoy later and post so more nuance can be gained from studying a particular set of users (viewers interested in economics and statistics videos may not display the same trends as those interested in popular culture). Overall, a YouTuber should think carefully about introducing ads where there were none playing before.



In general, we believe that this dataset is limited due to the following reasons: First, the videos are not popular enough to observe true differences between videos, and the statistical significance levels can be attributed more to the large sample size rather than differences in the success of videos. In addition, having low R-squared values across all analysis is concerning because it does not provide insight to where the majority of the variability is coming from for daily earnings and average watch time. There are a lot of other factors that should be explored here. Regardless of the R-square, the significant coefficients still represent the mean change in the dependent variable for one unit of change in predictor. However, notice that the R-square is so low that we won't be able to make any precise prediction, but can predict general pattern.

Additionally, one possible confounding factor in the design of the experiment is the unequal sample size across comparison group. When analyzing the effect of the post, pre-period group contains approximately 7000 more samples than post-period, which may be a potential confounding factor.

Despite these limitations in the data, we can still use the insights from this data for help in designing future studies and making limited connections between variables that YouTubers will be interested in.